**Data Ingestion:**

In any Enterprise Data Management, the core capability of the data lake architecture is the ability to quickly and easily ingest multiple types of data which can include Real-time streaming data and bulk data assets, Structured data generated and processed by legacy on-premises platforms.

Amazon Kinesis Data Firehose is part of the Kinesis family of services that makes it easy to collect, process, and analyze real-time streaming data at any scale. Kinesis Data Firehose is a fully managed service for delivering real-time streaming data directly to data lakes (Amazon S3), data stores, and analytical services for further processing. Kinesis Data Firehose automatically scales to match the volume and throughput of streaming data, and requires no ongoing administration.

Kinesis Data Firehose can also be configured to transform streaming data before it’s stored in a data lake built on Amazon S3. Its transformation capabilities include compression, encryption, data batching, minimizing the amount of storage used at the destination and increasing security. It can also transform the source data using AWS Lambda and deliver the transformed data to destinations. You configure your data producers to send data to Kinesis Data Firehose, which automatically delivers the data to the destination that you specify.

**Data enrichment**

When designing a streaming workload, it can sometimes be necessary to enrich the incoming data with external information. As data comes in, it might be saved into a raw datastore, but downstream consumers of this data want to use it to compute things like transactional fraud, application usage trends, stock purchase trends, and more. To calculate these important metrics, the stream must be enriched with external data.

Amazon Kinesis Data Streams enables you to build custom, real-time applications using popular stream processing frameworks and load streaming data into many different data stores which can perform data enrichment like looking up data from DynamoDB, and then produce the enriched data onto another stream.

In the given whitepaper, for “Internet offering based on location” The user details and location are enriched with different bandwidth options prior to publishing back to the application by integrating Amazon Kinesis data stream with AWS Lambda that enables this real time enrichment.

**Data processing**

Data processing occurs when data is collected and translated into usable information. Importance of data processing includes **increased productivity and profits, better decisions**, more accurate and reliable. Stream Data processing applications process data continuously in real-time, even before it is stored. Stream data processing platforms have to be able to handle the speed and variability of incoming data and process it as it arrives, often millions to hundreds of millions of events per hour.

When implementing a solution with Kinesis Data Streams, we can create custom data-processing applications known as Kinesis Data Streams applications. A typical Kinesis Data Streams application reads data from a Kinesis stream as data records. Data put into Kinesis Data Streams is ensured to be highly available and elastic, and is available in milliseconds.

**Data storage**

**Consumers (such as a custom application running on Amazon EC2 or an Amazon Kinesis Data Firehose delivery stream) can store their results using an AWS service such as Amazon DynamoDB, Amazon Redshift, or Amazon S3.**

A Kinesis data stream is an ordered sequence of data records meant to be written to and read from in real time. Data records are therefore stored in **shards in your stream temporarily**. The time period from when a record is added to when it is no longer accessible is called the retention period.

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance. S3 allows customers to store and protect any amount of data for virtually any use case, such as data lakes, cloud-native applications, and mobile apps. With cost-effective storage classes and easy-to-use management features.

Amazon Redshift is a fully-managed petabyte-scale cloud based data warehouse product designed for large scale data set storage and analysis. It is also used to perform large scale database migrations.

Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multi-Region, multi-active, durable database with built-in security, backup and restore, and in-memory caching for internet-scale applications. DynamoDB can handle more than ten trillion requests per day, and can support peaks of more than 20 million requests per second.

**Processed data delivery**

**Data delivery to Amazon S3**

**Data delivery to Amazon Redshift**

**Data delivery to Amazon ES**

**Data delivery to custom HTTP endpoints**

What are the different ways by which data can be written into and read from AWS Kinesis Streams? [1.75]

**Sending data to Amazon Kinesis Data Streams**

**Processing data in Amazon Kinesis Data Streams**

**Processing streams of data with AWS Lambda**

To read and process data from Kinesis streams, you need to create a consumer application. There are varied ways to create consumers for Kinesis Data Streams. Some of these approaches include using Amazon Kinesis Data Analytics to analyze streaming data using KCL, using AWS Lambda, AWS Glue streaming ETL jobs, and using the Kinesis Data Streams API directly

Q3. Briefly explain in your understanding of the situation (excluding given scenarios) where you will consider using Kinesis Streams over Firehose.

[**https://medium.com/aws-architech/use-case-kinesis-data-streams-vs-kinesis-data-firehose-74d639214e89**](https://medium.com/aws-architech/use-case-kinesis-data-streams-vs-kinesis-data-firehose-74d639214e89)

[**https://jayendrapatil.com/aws-kinesis-data-streams-vs-kinesis-firehose/**](https://jayendrapatil.com/aws-kinesis-data-streams-vs-kinesis-firehose/)

What are the ways by which streaming data can be queried using SQL?

Pg – 24 :

With Amazon Kinesis Data Analytics, you can interactively query streaming data using multiple options, including Standard SQL, Apache Flink applications in Java, Python and Scala, and build Apache Beam applications using Java to analyze data streams.

**Pg – 25**

As part of Kinesis Data Analytics service, Kinesis Data Analytics Studio is available for customers to interactively query data streams in real time, and easily build and run stream processing applications using SQL, Python, and Scala. Studio notebooks are powered by Apache Zeppelin.

Q5. Design and draw a streaming data pipeline using AWS components satisfying the following requirements - [3]

- Ingested streaming data to be stored for 7 days -- S3 and can use Athena to query data in S3

- Raw streaming data to be stored so that it can be fed to other AWS services --- From S3, you can fed this raw data into any other AWS services like AWS Lambda, Glue, Athena, QuickSight, Redshift or RDS.

- Scala to be used as language for data enrichment application – Use of Amazon Kinesis Data Streams and Lambda to do the data enrichment while streaming the data into S3

- Data at rest to be queried using SQL –You can use Athena to query through S3 using Glue Data Catalog.

- has to have capabilities to search within streaming enriched data -- **Amazon Elasticsearch Service**

- alert to be sent to the consumers when certain criteria is meet -- Amazon Kinesis Data Streams then invokes AWS Lambda functions, which can send the alerts to the driver and the fleet monitoring team through Amazon SNS.